Swin SMT: Global Sequential Modeling for Enhancing 3D Medical Image Segmentation

Szymon Płotka[1,2], Maciej Chrabaszcz[2], and Przemyslaw Biecek[2]

[1] Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands
[2] Warsaw University of Technology, Warsaw, Poland
maciej.chrabaszcz.dokt@pw.edu.pl

Abstract. Recent advances in Vision Transformers (ViTs) have significantly enhanced medical image segmentation by facilitating the learning of global relationships. However, these methods face a notable challenge in capturing diverse local and global long-range sequential feature representations, particularly evident in whole-body CT (WBCT) scans. To overcome this limitation, we introduce Swin Soft Mixture Transformer (Swin SMT), a novel architecture based on Swin UNETR. This model incorporates a Soft Mixture-of-Experts (Soft MoE) to effectively handle complex and diverse long-range dependencies. The use of Soft MoE allows for scaling up model parameters maintaining a balance between computational complexity and segmentation performance in both training and inference modes. We evaluate Swin SMT on the publicly available TotalSegmentator-V2 dataset, which includes 117 major anatomical structures in WBCT images. Comprehensive experimental results demonstrate that Swin SMT outperforms several state-of-the-art methods in 3D anatomical structure segmentation, achieving an average Dice Similarity Coefficient of 85.09%. The code and pre-trained weights of Swin SMT are publicly available at https://github.com/MI2DataLab/SwinSMT.

Keywords: 3D Image Segmentation · Sequential Modeling · Vision Transformer

1 Introduction

Segmentation of anatomical structures is crucial for various clinical applications such as diagnostics, treatment planning, quantitative analysis, image-guided interventions, and clinical trials [12]. In recent years, various segmentation methods of anatomical structures have been developed [3,4], particularly for specific tasks such as organ segmentation and tumor delineation [1,6], as well as for specific body parts like abdominal organs [8,15,16].

While large-scale ViT-based methods have garnered significant attention for their impressive performance and capacity to model long-range sequences, they

S. Płotka and M. Chrabaszcz – Authors contributed equally.

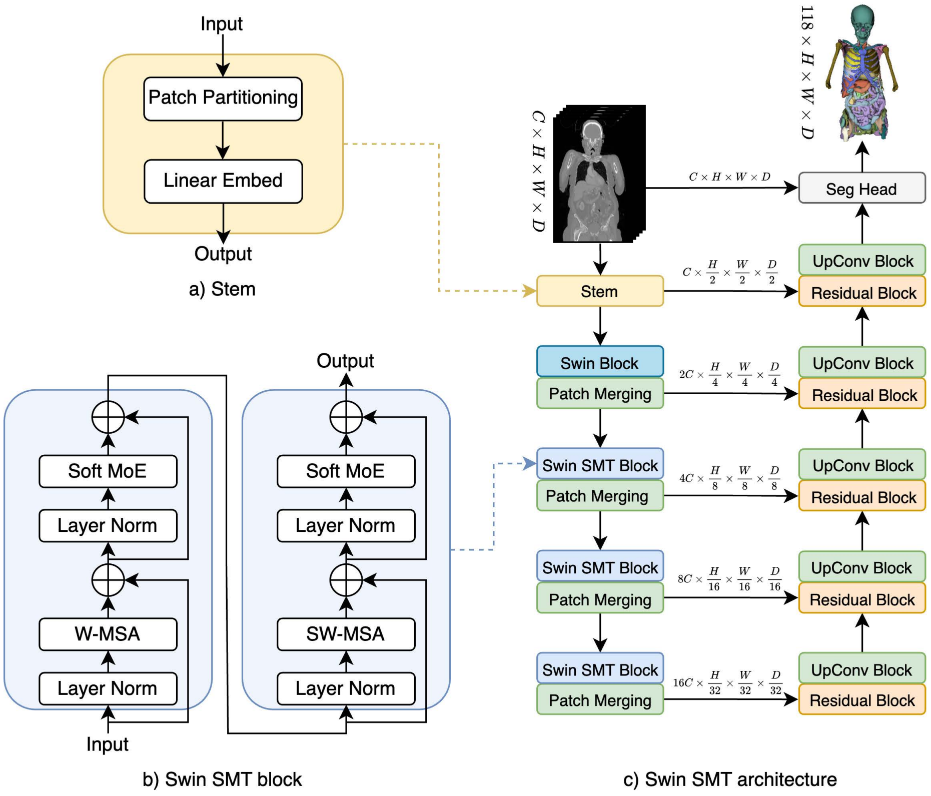

Fig. 1. An overview of the Swin SMT architecture. The input to our model is a 3D CT scan. The Stem creates non-overlapping patches of the input data and utilizes a patch partition layer to generate windows with a desired size for computing self-attention. Encoded feature representations in each of the encoder blocks are then fed to a Convolutional Neural Network (CNN)-based decoder via skip connections at multiple resolutions. The segmentation output consists of 118 channels, corresponding to 117 classes and background, representing the major anatomical structures in WBCT images. H, W, D, and C refer to height, width, depth, and number of feature channels, respectively. W-MSA and SW-MSA refer to window-based multi-head self-attention with regular and shifted windows, respectively.

encounter notable challenges when employed in medical image analysis. These challenges include computational complexity, as seen in architectures such as ViT-based backbone (e.g. UNETR [3]), RepUX-Net [10], or nnFormer [22], and a deficiency in modeling long-range dependencies in scale and intensity, as observed in models like Swin UNETR [19] or UNesT [21].

In recent years, Mixture-of-Experts (MoE) [11] has gained popularity in the Natural Language Processing domain and has subsequently been applied in the computer vision domain. MoE allows the model to dynamically allocate computational resources to the most relevant parts of the input data, enhancing efficiency and performance. By selectively activating different subset of the model's parameters, MoE enables the ViT to scale up its capacity without a corresponding increase in computational costs, leading to more accurate and efficient processing of complex visual information [11,18]. Unlike vanilla MoE, recently introduced Soft MoE [18] aggregates sequences into multiple global representations by using a weighted average over sequence elements, enabling the model to select the most important features. In the case of WBCT scans, global representations are crucial for the proper segmentation of certain classes with various scales and intensities, such as ribs and vertebrae.

Following these findings, we propose a novel **Swin S**oft **M**ixture **T**ransformer (Swin SMT), based on the Swin UNETR architecture, that incorporates Soft MoE to effectively handle complex and diverse long-range dependencies, as seen in WBCT images. Soft MoE allows for scaling up the model parameters maintaining a balance between computational complexity and segmentation performance in both training and inference modes. Swin SMT is designed for modeling the whole volume global features at various scales and intensities. To the best of our knowledge, this marks the first attempt to employ Soft MoE for the segmentation task. It is also the first benchmark of the publicly available TotalSegmentator-V2 dataset [20], encompassing 117 major anatomical structures in WBCT scans.

2 Method

Our Swin SMT is built upon the Swin UNETR [19], with a particular focus on enhancing the Transformer encoder. We incorporate the Soft MoE by replacing the feedforward network (FFN) within the Swin ViT block. The CNN-based decoder remains consistent with the original implementation [19]. An overview of the proposed method is illustrated in Fig. 1.

2.1 Swin Transformer Block

Let the input $X \in \mathbb{R}^{C \times H \times W \times D}$ be a 3D CT scan. Then, Stem layer first partitions scan X into a sequence of $2 \times 2 \times 2$ scan patches returning $X' \in \mathbb{R}^{(HWD/8) \times 8C}$ and later applies a linear layer. With input X, the stem layer can be described as:

$$X' = \text{Partition}(X), \quad X_{\text{out}} = \text{Linear}(X'), \tag{1}$$

where $X_{out} \in \mathbb{R}^{(HWD/8) \times d}$, and d is the hidden size.

We use the Swin block at the first stage due to its reported computational and performance efficiency [19]. While Swin blocks excel at creating local representations, they encounter issues with long-range dependencies due to the windowed attention mechanism used in them, a challenge we aim to mitigate in our model with later stages.

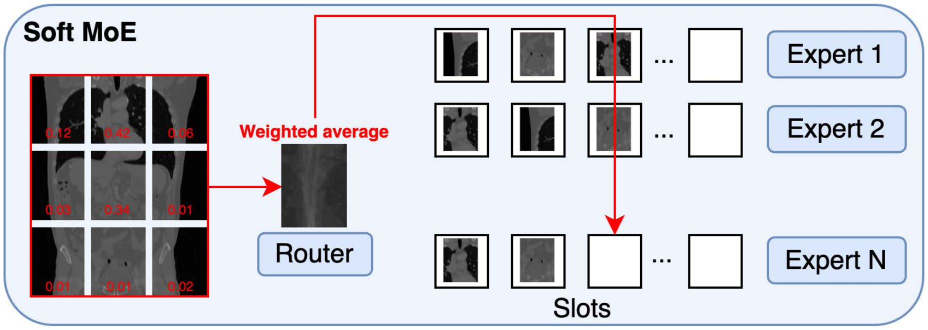

Fig. 2. An overview of the Soft MoE. Here, the router assigns the weighted average of all the input tokens (patches) to each slot, which computes logits for each input pair of tokens and slots using dispatch weights. Then, each expert processes its slots. Finally, the original logits are normalized per token and used to combine all the slot outputs for every input token. Tokens in slots are shown in decreasing order of logits assigned to this token by dispatch weights.

2.2 Swin Soft Mixture Transformer Block

Soft MoE calculates representations as convex combinations of tokens. This allows our model to better capture global information about WBCT scans by enabling each slot to create a comprehensive aggregation of input tokens. An overview of the Swin SMT block is presented in Fig. 2.

Motivated by [18], we swap FFN with Soft MoE layers within the Swin Transformer block. Due to memory complexity, Soft MoE layers are added in each stage, except for the first stage. Let $X \in \mathbb{R}^{m \times d}$ denote the input sequence, where m is sequence length, and d is hidden dimension size. With the model having n experts and $s = m/n$ slots per expert, the memory complexity of Soft MoE is $\mathcal{O}(m^2 + md + nd^2)$. For the Swin block, we have $m = (p/2^i)^3$ at each stage, where p is the patch size, and i is the stage number ($i \in 1, 2, 3, 4$). Then, for $p = 128$ at the first stage, we have $m = 262,144$, leading to significant memory demand. Therefore, we choose to skip adding Soft MoE at the first stage.

The Soft MoE layer computes slot representations as convex combinations of input tokens and learnable parameters Φ. The slot representations $\tilde{X}$ are computed according to the following equations:

$$D_{ij} = \frac{\exp((X\Phi)_{ij})}{\sum_{i'=1}^{m} \exp((X\Phi)_{i'j})}, \qquad \tilde{X} = D^T X. \tag{2}$$

After obtaining the slot representations $\tilde{X}$, n expert functions are applied to each of the corresponding s slots. The expert functions are modeled as FFNs that transform $\tilde{X}$. The transformation is expressed as:

$$Y = concat(\{FFN_{\lfloor i/n \rfloor}(\tilde{X}_i); i \in \{1, \ldots, ns\}\}), \tag{3}$$

where Y is the output matrix obtained by concatenating the results of the n expert functions, $FFN_{\lfloor i/n \rfloor}(\tilde{X}_i)$ represents the output of the $\lfloor i/n \rfloor$-th expert function applied to the i-th slot representation $\tilde{X}_i$, concat$(\{\ldots\})$ signifies the concatenation of the outputs from all expert functions.

Having calculated Y, we now compute dispatch weights, creating a convex combination of expert representations Y as follows:

$$S_{ij} = \frac{\exp\left((X\varPhi)_{ij}\right)}{\sum_{j'=1}^{n\cdot s} \exp((X\varPhi)_{ij'})}, \quad X_{out} = SY. \tag{4}$$

Here, S_{ij} represents the dispatch weight for the i-th input sequence token and the j-th slot representation, and X_{out} is the final output, obtained by multiplying the dispatch weights matrix S with the expert representation matrix Y.

2.3 Swin Soft Mixture Transformer

Encoder. Our encoder module is based on Swin Transformer [14] architecture with Soft MoE layers replacing FFN layers in Swin blocks at stages 2, 3, and 4. Having the input of previous layer $z_{l-1} \in \mathbb{R}^{m \times d}$, the outputs for those stages are calculated as follows

$$\begin{aligned} \hat{z}_l &= \text{W-MSA}(\text{LN}(z_{l-1})) + z_{l-1}, \quad z_l = \text{Soft-MoE}(\text{LN}(\hat{z}_l)) + \hat{z}_l \\ \hat{z}_{l+1} &= \text{SW-MSA}(\text{LN}(z_l)) + z_l, \quad z_{l+1} = \text{Soft-MoE}(\text{LN}(\hat{z}_{l+1})) + \hat{z}_{l+1}, \end{aligned} \tag{5}$$

where LN refers to the layer-norm layer, W-MSA, and SW-MSA represent window-based multi-head self-attention with regular and shifted windows, respectively, following the original Swin ViT implementation.

Decoder. Our decoder module is based on the same architecture as the decoder in the Swin UNETR model. Convolutional blocks are employed to extract outputs from each of the four blocks and the bottleneck. These extracted features are then upsampled using deconvolutional layers and concatenated with features from a higher resolution level. Finally, convolution with a 1 × 1 × 1 kernel is applied to map features to segmentation maps.

3 Experiments and Results

In this section, we describe the dataset used for training and evaluation of the Swin SMT. Additionally, we provide implementation details along with qualitative and quantitative results. Through an ablation study, we demonstrate the importance of the Swin SMT component that has been integrated into the Swin Transformer block.

Dataset. To train and evaluate our method, we use the publicly available TotalSegmentator-V2 dataset [20]. The dataset contains 1,228 CT scans with annotations for 117 major anatomical structures in WBCT images. It includes scans of the WBCT, as well as thoracic, abdominal, neck, and pelvic scans. Additionally, the dataset contains clipped sub-volumes corresponding to these body

Table 1. Quantitative results comparing Swin SMT with several state-of-the-art methods on the test set of TotalSegmentator-V2 are presented. We compare model parameters (in millions, M), inference time (in seconds, s), and segmentation performance results using DSC (%). Due to limited space, we divided the 117 classes into subgroups: organs, vertebrae, muscles, ribs, vessels, and the average of all these classes. The best-performing results are highlighted in bold. (*) denotes statistically significant differences between Swin SMT and compared state-of-the-art methods.

Method	**Params (M)**	**Time (s)**	**Organs** ↑	**Vertebrae** ↑	**Muscles** ↑	**Ribs** ↑	**Vessels** ↑	**Overall** ↑
UNETR [3]	102.02	35	73.84	60.70	82.35	69.27	61.49	70.88 (*)
SwinUNETR-S [19]	18.34	15	78.21	63.43	85.02	69.98	62.23	73.90 (*)
nnFormer [22]	149.30	99	79.26	73.87	74.97	74.03	74.97	75.48 (*)
DiNTS [5]	147.00	150	80.05	71.42	85.32	73.71	70.13	77.64 (*)
UNesT [21]	87.30	45	80.75	71.93	86.43	72.79	69.61	77.70 (*)
3D UX-Net [9]	53.00	74	83.03	79.54	86.99	82.54	75.01	82.53 (*)
SwinUNETR-B [19]	72.76	37	83.46	79.76	87.57	82.61	75.23	82.81 (*)
nnU-Net [7]	370.74	300	82.02	82.89	86.98	85.27	75.51	83.44 (*)
SwinUNETR-L [19]	290.40	145	83.26	82.02	87.99	83.82	75.60	83.59 (*)
3D RepUX-Net [10]	65.80	80	80.85	84.00	87.63	84.22	75.91	83.81 (*)
Universal Model [13]	62.25	39	82.25	**84.46**	87.58	86.49	76.11	84.02 (*)
Swin SMT (ours)	170.78	60	**83.70**	83.03	**88.70**	**86.60**	**77.54**	**85.09**

parts. We resample all scans to 1.5 × 1.5 × 1.5 mm^3 isotropic resolution. We follow the original data split [20] that uses 1082, 57, and 89 cases for training, validation, and testing, respectively.

Implementation details. We randomly crop a patch size of 128 × 128 × 128 from around the input CT scans. Training is conducted using the batch size of 1 per GPU and the AdamW optimizer with a warm-up cosine scheduler for 1000 epochs. As a loss function L, we use a sum of Dice and Cross-entropy loss defined as $L = L_D + \Lambda L_{CE}$, where L_D, L_{CE} are Dice and Cross-entropy loss, respectively. A grid search optimization in the range [0.5, 1.0] was performed, which estimated optimal value $\Lambda = 1$. An initial learning rate is set to 1×10^{-4}, and the weight decay is set to 1×10^{-5}. We adopt data augmentations of intensity, rotation, and scaling on the fly during training. For training, we use a DGX workstation equipped with 8 × NVIDIA A100 40GB GPUs, utilizing the Distribution Data Parallel methodology. The intensity of the Hounsfield Unit was clipped to the range [-1024, 1024] and linearly scaled to [0,1]. During the inference, we use a sliding window with an overlap of 0.5 and a Gaussian kernel. To compare the inference times, we use a scan with the dimensions of [300, 300, 422]. We implement Swin SMT using Python 3.9, PyTorch 2.1 [17], and MONAI 1.2.0 [2]. We evaluate segmentation performance using the Dice Similarity Coefficient (DSC). We perform a one-way analysis of variance (ANOVA) to assess the significance of differences among the performance metrics of Swin SMT segmentation performance. We use $p < 0.05$ as a threshold for statistically significant differences.

Quantitative results. To assess the effectiveness and performance of Swin SMT, we conduct a comprehensive analysis with several state-of-the-art methods, i.e., nnU-Net [7], UNETR [3], 3D UX-Net [9], three configurations of Swin UNETR-based (S, B, and L) [19], 3D RepUX-Net [10], nnFormer [22], DiNTS [5],

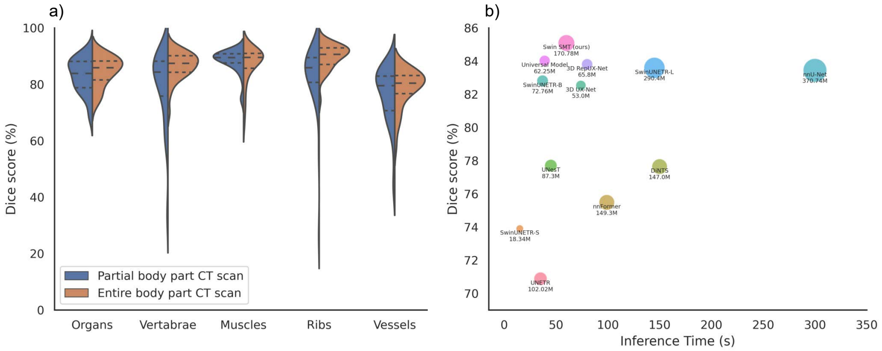

Fig. 3. a) Distribution of the quantitative results of Swin SMT for each subgroup within WBCT images. The partial and full body part CT scans denote cropped (i.e., a sub-volume of a thoracic or abdominal) and entire body part (i.e., thoracic, abdominal, or whole-body) CT scans, b) Distribution of the average DSC against inference time (in s). The inference time calculations are based on an input patch of 128 × 128 × 128 with a sliding window algorithm and overlay of 0.5. The size of each circle indicates the number of parameters (in M).

UNesT [21], and Universal Model [13] using the test set of TotalSegmentator-V2. To ensure a fair comparison, all models are trained under identical hardware settings. To enhance the performance parity among the compared methods, we employ the same hyperparameters, including optimizers, learning rates, and learning rate schedulers, following the configurations presented in their original works. Quantitative results for the TotalSegmentator's-V2 test set are presented in Table 1. Our proposed Swin SMT demonstrated superior segmentation performance across all subgroups except for vertebrae. Nevertheless, the Swin SMT achieves a higher average DSC of 85.09% across all 117 classes. During inference, our Swin SMT, with 170.78 million parameters, achieves a 60-second processing time. The one-way ANOVA revealed statistically significant differences between other state-of-the-art methods and Swin SMT ($p < 0.05$).

Ablation study. We conduct an ablation study to investigate the impact of varying the number of experts on the segmentation performance of the Swin SMT. We train five configurations as follows. First, we train the baseline method – Swin UNETR-B. Then, we swap FFN with Soft MoE in the Swin ViT block with 4, 8, 16, and 32 experts (n), respectively. Table 2 shows the results of an ablation study on the TotalSegmentator-V2 test set. The baseline method achieves a DSC of 82.81%. The incorporation of the Soft MoE ($n = 4$) into the baseline improved segmentation performance by 0.92%. Further increasing the number of experts to 2^n, where $n \in 3, 4, 5$ improved the DSC by 0.25, 0.56, and 1.36, respectively. The one-way ANOVA revealed statistically significant differences between the baseline and configurations with n experts, where n ranges from 4 to 32, including the best-performing configuration with $n = 32$

Table 2. The ablation study investigates the impact of varying the number of experts used in each Swin MoE block. We compare parameters (in millions, M), the number of experts (Experts), and DSC (%). (*) denotes statistically significant differences between the top-performing configuration with n = 32 experts and all other configurations.

Method	Params (M)	Experts (n)	DSC (%) ↑
Baseline	62.19	0	82.81 (*)
+ Soft MoE ($n = 4$)	83.62	4	83.73 (*)
+ Soft MoE ($n = 8$)	87.01	8	83.98 (*)
+ Soft MoE ($n = 16$)	114.93	16	84.29 (*)
+ Soft MoE ($n = 32$)	170.78	32	85.09

experts ($p < 0.05$). Due to limited computational resources, we did not use $> n = 32$ experts.

Performance analysis. Fig. 3(a) shows the distribution of the quantitative results of Swin SMT for each subgroup of the WBCT images. We compare the evaluation on the test set of TotalSegmentator-V2, where *Partial body part CT scan* and *Entire body part CT scan* denote cropped (i.e., a sub-volume of a thoracic or abdominal) and entire body part (i.e., thoracic, abdominal, whole-body) CT scans, respectively. We can observe that the Swin SMT exhibits robust segmentation performance on entire-body scans. However, when it comes to sub-volumes, the model lacks contextual information and shows low performance on context-related classes such as vertebrae, ribs, and vessels. In the supplementary material, we provide qualitative results for both partial and entire-body scans. We present both top-performing cases for each subgroup and qualitative errors observed in WBCT scans, even when the quantitative results per scan (DSC above 90%) are high.

Fig. 3(b) shows a speed-performance plot, indicating that Swin SMT presents a trade-off between computational efficiency and segmentation performance compared to several state-of-the-art methods while maintaining a high mean DSC score of 85.09%. Unlike Swin SMT, which is 1.54 times slower than the Universal Model (2nd best), it is 1.33, 2.42, and 5.0 times faster than the next three best-performing methods, including 3D RepUX-Net, Swin UNETR-L, and nnU-Net, respectively.

4 Conclusions

We proposed Swin SMT, a 3D Transformer-based method for segmenting major anatomical structures from WBCT images. By incorporating Soft MoE into the Swin Transformer block, Swin SMT effectively handles the complex and diverse long-range dependencies inherent in WBCT images. It is designed to model global features of the entire volume at various scales and intensities. A comprehensive benchmark, to the best of our knowledge, for the first time, on the publicly available TotalSegmentator-V2 dataset demonstrates superior performance compared to several state-of-the-art methods. Our method has the potential to serve as a support tool in clinical environments for the fast and accurate segmentation of major anatomical structures in WBCT images. In our future direction, we will incorporate large-scale self-supervised pretraining to enhance the performance of the segmentation model and evaluate its generalizability and transferability to external datasets.

Acknowledgments. This work was financially supported by the Polish National Center for Research and Development (grant no. INFOSTRATEG-I/0022/2021-00).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al.: The liver tumor segmentation benchmark (lits). Medical Image Analysis **84**, 102680 (2023)
2. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
3. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584 (2022)
4. He, Y., Nath, V., Yang, D., Tang, Y., Myronenko, A., Xu, D.: Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 416–426. Springer (2023)
5. He, Y., Yang, D., Roth, H., Zhao, C., Xu, D.: Dints: Differentiable neural network topology search for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5841–5850 (2021)
6. Heller, N., Isensee, F., Maier-Hein, K.H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., et al.: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. Medical Image Analysis p. 101821 (2020)
7. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
8. Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in Neural Information Processing Systems **35**, 36722–36732 (2022)
9. Lee, H.H., Bao, S., Huo, Y., Landman, B.A.: 3d UX-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. In: The Eleventh International Conference on Learning Representations (2023)
10. Lee, H.H., Liu, Q., Bao, S., Yang, Q., Yu, X., Cai, L.Y., Li, T.Z., Huo, Y., Koutsoukos, X., Landman, B.A.: Scaling up 3d kernels with bayesian frequency re-parameterization for medical image segmentation. In: International Conference

on Medical Image Computing and Computer-Assisted Intervention. pp. 632–641. Springer (2023)

11. Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., Chen, Z.: Gshard: Scaling giant models with conditional computation and automatic sharding. In: International Conference on Learning Representations (2020)
12. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017)
13. Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: Clip-driven universal model for organ segmentation and tumor detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21152–21164 (2023)
14. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
15. Luo, X., Liao, W., Xiao, J., Chen, J., Song, T., Zhang, X., Li, K., Metaxas, D., Wang, G., Zhang, S.: Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. Medical Image Analysis **82**, 102642–102642 (2022)
16. Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., et al.: Abdomenct-1k: Is abdominal organ segmentation a solved problem? IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6695–6714 (2021)
17. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems **32** (2019)
18. Puigcerver, J., Ruiz, C.R., Mustafa, B., Houlsby, N.: From sparse to soft mixtures of experts. In: The Twelfth International Conference on Learning Representations (2023)
19. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20730–20740 (2022)
20. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5) (2023)
21. Yu, X., Yang, Q., Zhou, Y., Cai, L.Y., Gao, R., Lee, H.H., Li, T., Bao, S., Xu, Z., Lasko, T.A., et al.: Unest: local spatial representation learning with hierarchical transformer for efficient medical segmentation. Medical Image Analysis **90**, 102939 (2023)
22. Zhou, H.Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L., Yu, Y.: nnformer: volumetric medical image segmentation via a 3d transformer. IEEE Transactions on Image Processing (2023)